Internet Archive Comments in Response to the United Kingdom's Open Consultation on Copyright and AI

February 25, 2025

These comments are provided on behalf of the Internet Archive, a non-profit research library based in the United States. Like other libraries, we work to expand access to knowledge by collecting, archiving, and providing public access to a variety of physical and digital collections. While our physical headquarters are located in San Francisco, California, most of our patrons, including those from the UK, visit our collections online at archive.org.

Overview and General Principles

Copyright law has been adapting to disruptive technologies since its earliest days; however, the UK's regime is inadequate to meet the current technological moment. The AI Opportunities Action Plan declared an optimistic vision for a UK policy regime that would shape the development and use of AI technologies to drive economic growth, benefit working people by improving health care and education and how citizens interact with their government; and open up new opportunities rather than threatening traditional patterns of work.[1]

The government's proposal to create an exception for text and data mining ("TDM") that permits rights holders to control which materials can actually be used for such purposes, labelled as "Option 3" in this Copyright and AI consultation, is not ambitious enough for the reasons stated below. Instead, we urge the government to reconsider the benefits that "Option 2" would provide for the development of a homegrown Artificial Intelligence (AI) industry in the UK.

Overall Comments

Regulation of AI must be considered holistically–not solely through the narrow lens of copyright law. AI has the potential to disrupt many professions, not just those of individual creators or even those working in copyright-intensive industries, such as the news media or film industry. Responses to this disruption must therefore be developed on an economy-wide basis. Further, copyright law should not be treated as a means for addressing broader societal challenges. Going down a typical copyright path of creating new rights and relicensing markets for AI, could serve to worsen social problems like inequality, surveillance, constraints on access to knowledge, and the monopolistic behaviour of technology firms and publishers.[2]

---

[1] Clifford, Matt, AI Opportunities Action Plan, Presented to Parliament by the Secretary of State for Science, Innovation and Technology by Command of His Majesty (January 2025).
[2] Craig, Carys J., The AI-Copyright Trap (July 15, 2024). Available at SSRN: https://ssrn.com/abstract=4905118 ("Copyright law is neither apposite nor equipped to govern the way that generative AI is developed, trained, deployed, or enjoyed. Insisting that it should do so, and imagining that it is up to the task, could do far more harm than good.")

Moreover, AI has the potential to impact nearly all sectors of the economy including healthcare, climate, energy, transportation, and education—not to mention potential military and other national security applications. Narrowly focusing on copyright law and over-indexing on potential negative impacts on a narrow band of copyright holders from the creative industries will therefore have serious consequences in these other areas.

Additionally, universities, libraries, and other publicly-oriented institutions must be able to continue to ensure the public's access to high-quality, verifiable sources of news, scientific research, and other information essential to their participation in society. Strong libraries and educational institutions can help mitigate some of the challenges to our information ecosystem, including those posed by AI. These mission-oriented institutions should be empowered to provide access to educational resources of all sorts–including the powerful AI tools now being developed.

If the UK wishes to not only compete on AI, but to become a global leader, then the best path forward is to implement a broad exception for TDM—Option 2 as described in this Consultation. Other approaches effectively allow publishing industries to act as gatekeepers to AI innovation; this will hold the UK back in ways that will make it impossible to compete with jurisdictions that have no such restrictions. In particular, US researchers far outpace those from other countries in publishing impactful papers based on TDM techniques, and US firms appear far ahead in the market. China appears to be the other global leader. This appears likely to remain the status quo without innovation-friendly policy interventions by other governments. Option 3 is not such an intervention, and will likely make the UK's standing in AI markets worse.

Responses to Specific Questions

Question 1. Do you agree that option 3 is most likely to meet the objectives set out above?

Whether or not Option 3 might meet the objectives laid out in this consultation, it will not serve to meet the objectives set forth in the AI Opportunities Action Plan. It would not lead to the UK becoming a global leader.

Question 2. Which option do you prefer and why?

Option 2 is preferable to the other options. Having a clear exception allowing TDM for AI training purposes would incentivise researchers, developers, and companies to choose the UK to do such work.[3] Options 1 and 3 require permission and payment for so-called "non-expressive" computational research activities. This would expand the scope of rights granted under copyright, allowing copyright holders to extract payment for underlying facts and information that

[3] *See, e.g.,* Martens, Bertin, Economic Arguments in Favour of Reducing Copyright Protection for Generative AI Inputs and Outputs (Bruegel, Working Paper Issue 09/2024, Apr. 4, 2024), available at https://www.bruegel.org/system/files/2024-04/WP%2009%20040424%20Copyright%20final_0.pdf

have nothing to do with creative expression.[4]

Option 2 could lead to legitimate use of datasets from cultural heritage organizations that could propel the UK to the forefront of AI development. For instance, the Internet Archive, Google, and others in the United States, have digital cultural heritage materials widely pursued as datasets for US based AI organizations. If there were regulatory clarity in the UK around the use of such materials, as Option 2 would provide, then it would be safe for researchers and innovators in the UK to do so.

The EU scheme, while seeking to provide such clarity, has in practice been confusing to many cultural heritage organizations and research organizations, leading to delays and possible cancellations of projects–if they even get off the ground in the first place. The Swiss "technology neutral" approach has proved, in practice, somewhat better. What this teaches is that, if the UK seeks to benefit from and lead in AI, regulatory clarity will be paramount.

Question 3. Do you support the introduction of an exception along the lines outlined above?

No.

Question 4. If so, what aspects do you consider to be the most important? If not, what other approach do you propose and how would that achieve the intended balance of objectives?

Creating a clear exception for TDM would best meet the UK's objectives in becoming a global leader in AI. And this need not be done at the expense of the holders of copyrights: protection of copyright holder interests will naturally come in the form of enforcement of those rights as against the particular *outputs* of AI systems that are themselves infringing.

Question 5. What influence, positive or negative, would the introduction of an exception along these lines have on you or your organisation? Please provide quantitative information where possible.

A clear TDM exception along the lines of Option 2 could empower libraries like ours to partner with the UK government, universities, and other mission-aligned organisations to work on public-interest AI tools that could benefit society as a whole.[5] Currently, access to digital

[4] *See, e.g.,* Sag, Matthew, The New Legal Landscape for Text Mining and Machine Learning (February 27, 2020), Journal of the Copyright Society of the USA, Vol. 66 p.291 (2019). Available at http://dx.doi.org/10.2139/ssrn.3331606 (Clarifying the difference between expressive and non-expressive uses of copyrighted works and arguing why allowing TDM and other similar non-expressive uses of copyrighted works without authorization is entirely consistent with the fundamental structure of copyright law). For additional challenges to the Option 3 approach see
Margoni, Thomas, TDM and Generative AI: Lawful Access and opt-outs (May 30, 2024). Forthcoming in Auterus&Media 2024. Available at SSRN: https://ssrn.com/abstract=5036164.

[5] OECD, "Enhancing Access to and Sharing of Data in the Age of Artificial Intelligence", available at https://www.oecd.org/en/publications/enhancing-access-to-and-sharing-of-data-in-the-age-of-artificial-intel

collections and the computational resources needed to utilise modern AI tools is centralised in a handful of for-profit companies.[6] Restrictive relicensing regimes will only further centralise their power.

Switzerland, for instance, has not directly regulated AI, but has opted so far for a "technologically neutral" approach to regulation. This has attracted thousands of workers from large tech firms (Google, Meta) and many innovative startups in Zurich. This regulatory framework is also attracting innovators and datasets from around the world.

The EU's text and data mining rules, while not without their challenges, are also leading cultural heritage organizations, research organisations, and entrepreneurs to support AI development there. Article 3 alone will likely lead to petabytes of data that might not be available in places without sufficient regulatory certainty. Without the UK providing at least as much legal certainty as Article 3, it will miss out on these opportunities.

Question 9. Is there a need for greater standardisation of rights reservation protocols?

Among the many upsides of Option 2 is the fact that there would be no need to develop rights reservation protocols at all.

Question 15. Should the government have a role in encouraging collective licensing and/or data aggregation services? If so, what role should it play?

The government should not encourage collective licensing in the AI context.[7] Collective licensing often benefits the biggest players in the content industry, while hindering innovation. Individual creators would likely see very little value from such a scheme. In the case of AI training—which requires such massive amounts of data—any individual contributions, and therefore remuneration, will likely be trivial.[8] Given that as a rule very limited digital rights are afforded UK collecting societies in the sectors where they exist, (there is effectively no collecting society for moving image in the UK), it is an opt-in scheme, and new licences take many years

---

ligence_23a70dca-en/full-report.html "OECD studies indicate that enhancing access to and sharing of both public and private-sector data can help unlock significant social and economic benefits, potentially contributing between 1% and 2.5% of GDP."

[6] *E.g., Widder, David Gray and West, Sarah and Whittaker, Meredith, Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI (August 17, 2023). Available at SSRN: https://ssrn.com/abstract=4543807 or http://dx.doi.org/10.2139/ssrn.4543807*

[7] Quintais, João Pedro, Generative AI, Copyright and the AI Act (January 30, 2025). Computer Law & Security Review, Volume 56, 2025, Available at SSRN: https://ssrn.com/abstract=4912701 "(… collective licensing faces significant substantive and practical challenges that need to be addressed to make it a viable option in this area.")

[8] Indeed, Spain abandoned plans to do collective licensing due to the criticisms of artists, writers and other cultural professionals, as well as from their respective organisations and unions. "Cultura retira su decreto para regular la IA ante la falta de consenso y abre un diálogo con los creadores" [Culture Ministry Withdraws Its Decree to Regulate AI Due to Lack of Consensus and Opens Dialogue with Creators], EL PAÍS (Jan. 29, 2025), available at https://elpais.com/cultura/2025-01-29/cultura-retira-su-decreto-para-regular-la-ia-ante-la-falta-de-consenso-y-abre-un-dialogo-con-los-creadores.html

to develop. This slow and highly piecemeal approach would not support the innovation outlined in the AI Opportunities Action Plan.[9]

Creation of a collective licensing scheme for AI training would therefore best serve the interests of very large corporate rightsholders, allowing them to become the gatekeepers over what information can be used to train AI models. This would have at least two negative effects. The first is to entrench only the largest tech companies who can afford to pay license fees, at the expense of smaller developers and likely even non-commercial AI projects. The second negative impact of collective licensing for AI training is that it would allow copyright holders to expand the scope of their copyrights well beyond the expressive elements of their work. Copyright is meant to encourage learning and research, not stifle the very creativity and innovation it is supposed to protect.

Question 17. Do you agree that AI developers should disclose the sources of their training material?

Internet Archive recommends that AI developers be encouraged to disclose training sources, but not legally mandated to do so. In the United States, there is no federal law requiring such transparency, and although there have been a variety of State law efforts to mandate transparency in various contexts, such laws are likely invalid.[10] As such, legal transparency mandates required by the UK government may well discourage developers from training in the UK when they could more easily do so in the United States.

Nevertheless, finding ways to encourage transparency could have major benefits. As a library, the Internet Archive seeks to support information integrity and authenticity. In the digital realm, this is often realised through metadata practices. Ideally these provide enough context and history for individuals to understand, in a verifiable manner, who produced or created a particular item, where it came from, and whether that item has been changed over time.

The Wayback Machine, our historical archive of public websites dating back to 1996, can serve as a useful example here. When we collect the content of a particular web page, we also collect provenance details. This practice allows researchers to see specific details of each snapshot, such as what entity collected the page, the URL and the time and date stamp of when it was collected. Including provenance information allows Wayback links to be used as a trusted citation in academic or journalistic work, by fact-checkers seeking to verify who said what on the web, and even in court. In this way, provenance helps to support a healthy information ecosystem by supporting the truth-seeking function of core democratic institutions while allowing ordinary readers to verify the source of a particular claim or assertion.

---

[9] We observe that UK collecting societies could have developed licences for machine learning since the 1960s when the technology took off, or again since 2014, but no licence for commercial AI exists in the UK.

[10] *See, e.g.*, Reporters Committee for Freedom of the Press, 9th Circuit: Provisions of California's content-moderation law violate First Amendment (Sept. 11 2024), available at https://www.rcfp.org/x-v-bonta-ninth-circuit-ruling/, and *X Corp v. Bonta*, 116 F. 4th 888 (9th Cir. 2024).

Presently, many AI chatbots and assistants are not built to show their work or give citations for the claims they make. When such chatbots produce answers to prompts, the source of the answers is generally not available because of the way those answers are generated. While there continue to be attempts to address this issue, provenance in the traditional sense–which has long been critical to knowledge creation and sharing–is not currently part of the AI ecosystem.[11]

While no doubt well meaning, a government transparency mandate that only has copyright in mind will not solve the provenance problem. Given the enormous size of the datasets used to train many LLMs, and the inability to credit individual answers to sources in those datasets, "transparency" will likely amount to undifferentiated lists of millions or billions of URLs. These lists would not give a researcher the ability to understand where any particular idea or concept came from when it is returned as a response to a question or prompt.

As the technology matures, and the issue of provenance continues to be addressed, copyright rules regarding transparency must be carefully crafted so as to further, and not to discourage, these efforts. The UK could, for example, offer a safe harbor for AI developers who do disclose their training sources in a manner that fosters a healthy information ecosystem.

Question 19. What transparency should be required in relation to web crawlers?

Since web crawling is among the tools used by AI companies to gather training materials, some lawmakers view regulating crawling as a good way to regulate those companies. Unfortunately, this approach is overly broad and likely to harm the public interest by preventing legitimate uses of publicly available data. Web crawling has been a cornerstone of how the open web has functioned for decades, and the UK government should be mindful not to disrupt the myriad socially beneficial uses of public web data gathered via crawlers.

Web crawling is the backbone of the Internet Archive's work of preserving access to public websites over time. Archiving public websites is vital for maintaining an informed citizenry, fostering transparency, and preserving historical records. Our historical web archive, the Wayback Machine, is used by journalists, researchers, lawyers, teachers, and other members of the general public on a daily basis for a wide range of purposes that benefit society. For example, we've been able to make Wikipedia more useful and reliable by restoring links to over 20 Million deleted web pages.

Beyond Internet Archive's activities, web crawling supports commercial uses such as the search engines that make the vast World Wide Web navigable by human beings and market research that is essential to companies and investors around the globe. Even more importantly, data gathered from the public web is a vital resource for academic and scientific research, offering a rich pool of information for exploratory studies and empirical research across all disciplines,

[11] *See, e.g.,* Cargnelutti, Matteo. *Did ChatGPT really say that?": Provenance in the age of Generative AI.* (May 22, 2023) Available at: https://lil.law.harvard.edu/blog/2023/05/22/provenance-in-the-age-of-generative-ai/

including health sciences, environmental studies, and technology. Public web data is also a necessity for non-profit organisations, governmental bodies, and academic institutions who rely on it for mission-driven research and projects. Among other uses, public web data can be used to understand and improve bias in the workplace, identify and report online hate speech or CSAM, track the spread of disease, develop policies to mitigate the dangers social media poses to today's youth, and more.

As such, regulating web crawling in a general way would impact many different types of users, from researchers to for-profit companies, and likely have unintended consequences. We suggest instead to focus more narrowly on regulating potentially harmful practices of AI companies.

Question 28. Does the existing data mining exception for non-commercial research remain fit for purpose?

If the government moves forward with its current proposal (Option 2), then we would urge that it also retain a non-commercial exception along the lines of the existing exception set forth in Section 29A of the UK Copyright, Designs and Patents Act. This approach would, at a minimum, put the UK on a similar footing as that found in the EU 2019 Copyright in the Digital Single Market Directive.

However, as written, the UK's existing exception is more burdensome than necessary. Take for one example a researcher seeking to use TDM methods on older physical materials such as scrolls or hand-written letters.To the extent that a researcher wishes to use such materials as part of her TDM project, then they must be digitised. That digitisation process is labour-intensive and expensive to undertake. A non-profit research organisation or educational institution seeking to digitise such materials should not be *prohibited from ever using those scans for any other purpose.* Similarly, the existing exception does not clearly support sharing amongst project partners, or remote access. Non-commercial actors, many of which are government-funded, should be encouraged to work together and be efficient, not to have to digitise the same materials over and over.

Moreover, Section 29A does not clearly enable public-private partnerships. These have proved quite beneficial elsewhere. For example, beginning in 2004 Google partnered with a number of large research libraries in order to scan millions of books that became the basis of the HathiTrust Digital Library (HDL).[12] This effort cost Google an estimated $400 million dollars, while it cost the libraries nothing at all. The partnership allowed Google to develop any number of products from its original "snippets" search feature to its new AI tool, Gemini. This same effort

[12] For far more detailed information about this project, see Marcum, Deanna and Roger C. Schonfeld, Along Came Google: A History of Library Digitization (Princeton University Press 2023).

supported HDL to develop its own researcher access programs,[13] enabling countless academic TDM projects over the last two decades.[14]

Finally, without amending the UK's out-dated circumvention of technological protection measures provisions, any TDM exceptions will be ineffective.

[13] https://www.hathitrust.org/the-collection/search-access/ and https://www.hathitrust.org/member-libraries/resources-for-librarians/data-resources/

[14] For just one example of the computational research techniques empowered by access to massive digital collections, *see*, Underwood, Ted, David Bamman, and Sabrina Lee. "The Transformation of Gender in English-Language Fiction." *Journal of Cultural Analytics,* Vol. 3, Issue 2 (Feb 2018). Available at https://culturalanalytics.org/article/11035.